SUPPLEMENTARY APPENDIX

The reliability of immunohistochemical analysis of the tumor microenvironment in follicular lymphoma: a validation study from the Lunenburg Lymphoma Biomarker Consortium

Birgitta Sander,[1] Daphne de Jong,[2] Andreas Rosenwald,[3] Wanling Xie,[4] Olga Balagué,[2] Maria Calaminici,[5] Joaquim Carreras,[6] Philippe Gaulard,[7] John Gribben,[5] Anton Hagenbeek,[8] Marie José Kersten,[8] Thierry Jo Molina,[9] Abigail Lee,[5] Santiago Montes-Moreno,[10] German Ott,[11] John Raemaekers,[12] Gilles Salles,[13] Laurie Sehn,[14] Christoph Thorns,[15] Björn E. Wahlin,[16] Randy D. Gascoyne,[14] and Edie Weller[4]

[1]Department of Laboratory Medicine, Division of Pathology, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden; [2]Department of Pathology, Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands; [3]Institute of Pathology, University of Würzburg, Würzburg, Germany; [4]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA; [5]Centre for Haemato-Oncology, Barts Cancer Institute, Queen Mary, University of London, London, UK; [6]Hematopathology Section, Department of Pathology, Hospital Clinic, IDIBAPS, University of Barcelona, Barcelona, Spain, *present address Tokai University, School of Medicine, Japan*; [7]Department of Pathology and Inserm U955, Hôpital Henri Mondor, University Paris-Est Créteil, France; [8]Academic Medical Center, Department of Hematology, Amsterdam, The Netherlands; [9]Université Paris-Descartes and AP-HP, Hôtel-Dieu, Paris, France; [10]Department of Pathology, Hospital Universitario Marqués de Valdecilla/IFIMAV, Santander, Spain; [11]Department of Clinical Pathology, Robert-Bosch-Krankenhaus, and Dr Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, Germany; [12]Department of Hematology, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands; [13]Hospices Civils de Lyon and Université Claude Bernard Lyon-1, UMR CNRS 5239, Lyon, France; [14]Department of Pathology and Medical Oncology, British Columbia Cancer Agency, University of British Columbia, Vancouver, Canada; [15]Department of Pathology, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany; and [16]Department of Medicine, Division of Hematology, Karolinska Institutet, Stockholm, Sweden

 doi:10.3324/haematol.2013.095257
Manuscript received on July 19, 2013. Manuscript accepted on December 19, 2013.
Correspondence: birgitta.sander@ki.se

Supplementary Materials and Methods

List of online supporting information

Supplementary Materials and Methods including statistical analysis

Table S1. Manual score distribution (N,%) across all labs and by lab for core 1.

Table S2. Manual score distribution (N,%) across all labs and by lab for core 2.

Table S3. Comparison of FOXP3, MIB1, CD68 scores from image analysis and manual scoring for core 1

Table S4. Summary statistics for the comparison of image and manual scoring for FOXP3, MIB1 and CD68 for core 1.

Table S5. Reasons for not scoring samples, by laboratory

Antibody panels for flow cytometry

Panel 1 (CD45-FITC, CD14-PE, CD19-PC5); Panel 2 (CD10-FITC, CD11c-PE, CD20-PECy5); Panel 3 (CD5-FITC, CD19-PE, CD3-PECy5); Panel 4 (CD7-FITC, CD4-PE, CD8-PECy5); Panel 5 (FMC7-FITC, CD23-PE, CD19-PECy5); and Panel 6 (Kappa-FITC, Lambda-PE, CD19-PECy5). All antibodies were from Beckman Coulter.

Criteria and scoring methods for immunohistochemistry

Pathologists reviewed the slides to exclude non-representative cores, the main reasons being lack of lymphoma tissue. For core 1, 1 case was excluded for CD3, CD4, CD8, CD68 and Ki67, 2 for CD21, CD34 and FOXP3 and 3 for CD10. For core 2, 1 case was excluded for CD3, CD4, CD8, CD68 and Ki67, 1 for CD21, 2 for FOXP3, 3 for CD34 and 4 for CD10. The number of cases included in the analysis ranged from 22-24 for core 1 (Table S1) and 21-24 for core 2 (Table S2).

Briefly, the same slide set was rotated between the scoring laboratories/pathologists and independently scored by each person, guided by an instruction manual with

representative images for comparison and without knowledge of the results of other scorers, flow cytometry data or image analysis. Scores were reported on an Excel data sheet and the results were centrally collected for analysis. CD10 was scored on tumor cells in neoplastic follicles as weak or strong positive staining versus negative. CD3, CD4, CD8 and FOXP3 positive T-cell populations were scored as percentages of all nucleated cells at 4 levels: 0-5%, 6-25%, 26-50%, >50% per total cell numbers in each core. Inter- and intrafollicular areas were not recorded separately. CD68 was counted as percentages at 5 levels 0-10%, 11-20%, 21-30%, 31-50%, >50%. The architectural patterns for CD3, CD4, CD8 and FOXP3 positive T-cell populations and for CD68 positive cells were scored in three categories as predominantly interfollicular, predominantly intrafollicular, and diffuse (=combined pattern). For FOXP3 the category perifollicular was also included. MIB1 was scored in follicles only, using the cut-offs 0-5%, 6-25%, 26-50% and >50%. CD21 staining was scored in follicles only and scored as well developed, partly disrupted but mostly intact, and mostly disrupted/absent. CD34 expressed by microvessels was scored as dense, moderately dense and sparse.

Scoring by automated microscopy was done as follows: For CD10, CD21 and MIB1, staining was assessed in follicles only, for all other markers, percentages were used of the total number of cells/core. The values are expressed as % of cells positive for a given marker/total numbers of cells in the core. Thus, the image analysis results in a continuous variable of positivity while the manual scoring was reported in categories.

Statistical analysis

Summary statistics are reported by laboratory and core for manual scores (counts and percentages) and for image and flow scores (median, mean, standard deviation, minimum, maximum). To test for manual score agreement by laboratory, analysis of

variance of the rank scores was performed across all markers (Wilk's lambda test, $p<0.0001$) and by marker (F-test) with Tukey adjustment for multiple comparison adjustment. In this analysis, the 7 laboratory scores are ranked for each patient and then averaged over the patient's scores by laboratory for each marker. Under perfect agreement, the average rank would be 4 for each laboratory. Values < 0.05 were considered statistically significant. To quantify agreement, the proportion of patients for whom all laboratories agree and all but one lab agrees is reported as well as the average pairwise agreement between laboratories. The average pair-wise agreement was adjusted for the expected proportion of agreement under the null model using the free marginal Kappa statistics of Brennan and Prediger[1,2], which is useful when the raters are not forced to assign a certain number of cases to each category. This statistic was selected because it minimizes the bias due to prevalence dependency [3] and is consistent with the scoring approach used. The level of agreement for the free marginal kappa statistic was evaluated using the following ranges: <0.40 low, 0.40-0.75 moderate and >0.75 high. The bootstrap method (2000 replicates) was used to estimate the standard error with the confidence intervals calculated based on the percentiles of the bootstrap distribution [4]. To preserve the correlation structure of the scores of each patient, the resampling was performed at the patient level. The agreement metrics were evaluated including the not scored category. This analysis is performed for the two cores, separately.

To compare the flow cytometric and image analysis score distribution to the manual scores, the distribution of the flow and image (continuous) scores are compared within the manual scoring categories using Jonckheere Terpstra test. Agreement is evaluated using misclassification measures to evaluate agreement between flow and image vs.

manual scoring, the continuous image and flow values are first categorized into the manual groups. Next, the proportion of the flow and image scores which agree within the manual category is determined, and if they disagree, the proportion which are under versus over classified is reported. Similar measures were computed to determine if the flow and image scores are within 5% of the manual category (that is, no more than 5% less than the lower cut-point of the manual category and no more than 5% higher than the upper cut-point for the manual category).

Spearman coefficient and the concordance coefficient are used to evaluate correlation and agreement between flow cytometric and image analysis, and the two image analysis machines [5].

1. Brennan RL, Predigar, D.J. Coefficient kappa: some uses, misuses, and alternatives. Educ Psychol Meas 1981;41:687-99.
2. Mackay A, Weigelt B, Grigoriadis A, Kreike B, Natrajan R, A'Hern R*, et al.* Microarray-based class discovery for molecular classification of breast cancer: analysis of interobserver agreement. J Natl Cancer Inst. 2011;103(8):662-73.
3. von Eye A, Mun EU. Analyzing Rater Agreement: Manifest Variable Methods. Lawrence Erlbaum Associates, Inc, . 2005;4:14-6.
4. Efron B, Tibshirani, R.J. An Introduction to the Bootstrap. New York, NY, Chapman & Hall. 1993.
5. Lin L. A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989;45(1):255–68.

Table S1. Manual score distribution (N,%) across all labs and by lab for Core 1. Exclude not scored

Marker/score	All labs	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7
CD10								
02=Positive/Weak	33(22)	5(24)	4(18)	5(23)	4(18)	5(23)	5(23)	5(25)
03=Positive/Strong	118(78)	16(76)	18(82)	17(77)	18(82)	17(77)	17(77)	15(75)
CD21								
01=Well developed	56(35)	6(26)	12(52)	7(30)	10(45)	14(61)	5(22)	2(9)
02=Partly disrupted but mostly intact	69(43)	11(48)	8(35)	9(39)	10(45)	5(22)	11(48)	15(68)
03=Mostly disrupted	27(17)	5(22)	2(9)	6(26)	1(5)	3(13)	6(26)	4(18)
04=Absent	7(4)	1(4)	1(4)	1(4)	1(5)	1(4)	1(4)	1(5)
CD3								
01=0-5%	3(2)					3(13)		
02=5-25%	42(26)	3(13)	5(21)	5(21)	7(32)	7(29)	6(25)	9(39)
03=25-50%	79(48)	4(17)	15(63)	15(63)	11(50)	11(46)	13(54)	10(43)
04=>50%	40(24)	16(70)	4(17)	4(17)	4(18)	3(13)	5(21)	4(17)
CD34								
01=Dense	14(9)	2(9)	2(9)	3(13)	1(5)	2(9)	2(9)	2(9)
02=Moderately dense	35(22)	4(17)	5(22)	4(17)	4(18)	1(4)	8(35)	9(41)
03=Sparse	110(69)	17(74)	16(70)	16(70)	17(77)	20(87)	13(57)	11(50)
CD4								
01=0-5%	5(3)					5(21)		
02=5-25%	60(37)	7(30)	6(25)	6(26)	8(38)	8(33)	10(42)	15(65)
03=25-50%	74(46)	11(48)	15(63)	12(52)	9(43)	9(38)	12(50)	6(26)
04=>50%	23(14)	5(22)	3(13)	5(22)	4(19)	2(8)	2(8)	2(9)
CD68_PCT								
01=0-10%	85(61)	9(39)	24(100)	19(79)	7(33)	8(35)		18(75)
02=10-20%	34(24)	8(35)		5(21)	9(43)	6(26)		6(25)
03=20-30%	14(10)	5(22)			5(24)	4(17)		
04=30-50%	6(4)	1(4)				5(22)		
CD8								
01=0-5%	36(22)		1(4)	4(17)	15(65)	4(17)	2(8)	10(43)
02=5-25%	112(68)	14(61)	21(88)	18(78)	7(30)	19(79)	21(88)	12(52)
03=25-50%	9(5)	8(35)	1(4)					
04=>50%	7(4)	1(4)	1(4)	1(4)	1(4)	1(4)	1(4)	1(4)
FOXP3_PCT,								
01=0-5%	95(60)	3(14)	13(57)	11(48)	17(74)	13(57)	18(78)	20(91)
02=5-25%	58(37)	16(76)	10(43)	12(52)	6(26)	7(30)	5(22)	2(9)
03=25-50%	5(3)	2(10)				3(13)		
Ki67								
01=0-5%	24(15)	2(10)	4(18)	1(4)	4(20)	10(45)		3(13)
02=5-25%	57(37)	14(67)	5(23)	9(39)	4(20)	10(45)	7(29)	8(35)
03=25-50%	59(38)	5(24)	10(45)	11(48)	9(45)	1(5)	14(58)	9(39)
04=>50%	15(10)		3(14)	2(9)	3(15)	1(5)	3(13)	3(13)
CD3_ARCH								
01=Intrafollicular	3(2)			1(4)		1(4)		1(4)
02=Interfollicular	119(72)	18(78)	17(71)	11(46)	18(75)	14(58)	21(88)	20(87)
03=Diffuse	44(27)	5(22)	7(29)	12(50)	6(25)	9(38)	3(13)	2(9)
CD4_ARCH								
01=Intrafollicular	8(5)	1(4)	1(4)	1(4)	1(5)	2(8)	1(4)	1(4)
02=Interfollicular	119(72)	18(75)	17(71)	11(48)	18(82)	14(58)	19(79)	22(92)
03=Diffuse	38(23)	5(21)	6(25)	11(48)	3(14)	8(33)	4(17)	1(4)
CD68_ARCH								
01=Intrafollicular	13(8)			4(17)		5(22)	4(17)	
02=Interfollicular	73(45)	11(48)	12(50)	10(42)	6(29)	9(39)	11(46)	14(61)
03=Diffuse	76(47)	12(52)	12(50)	10(42)	15(71)	9(39)	9(38)	9 (39)
CD8_ARCH								
01=Intrafollicular								
02=Interfollicular	143(86)	20(87)	19(79)	21(91)	21(88)	17(71)	22(92)	23(96)
03=Diffuse	23(14)	3(13)	5(21)	2(9)	3(13)	7(29)	2(8)	1(4)
FOXP3_ARCH								
01=Intrafollicular	10(6)	2(9)			1(4)	5(22)		2(9)
02=Interfollicular	61(38)	10(45)	9(39)	11(48)	5(22)	7(30)	13(57)	6(27)
03=Diffuse	64(40)	8(36)	10(43)	8(35)	11(48)	11(48)	6(26)	10(45)
04=Perifollicular	24(15)	2(9)	4(17)	4(17)	6(26)		4(17)	4(18)

Table S2. Manual score distribution (N,%) across all labs and by lab for Core 2. Exclude not scored

Marker/score	All Labs	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7
CD10								
02=Positive/Weak	28(19)	5(24)	3(14)	4(19)	4(19)	3(14)	4(19)	5(25)
03=Positive/Strong	118(81)	16(76)	18(86)	17(81)	17(81)	18(86)	17(81)	15(75)
CD21								
01=Well developed	53(32)	7(30)	11(46)	7(29)	9(41)	13(54)	3(13)	3(13)
02=Partly disrupted but mostly intact	78(48)	9(39)	11(46)	10(42)	11(50)	7(29)	14(58)	16(70)
03=Mostly disrupted	26(16)	6(26)	1(4)	6(25)	1(5)	3(13)	6(25)	3(13)
04=Absent	7(4)	1(4)	1(4)	1(4)	1(5)	1(4)	1(4)	1(4)
CD3								
01=0-5%	2(1)					2(9)		
02=5-25%	49(30)	4(17)	6(25)	7(29)	10(43)	6(26)	8(33)	8(33)
03=25-50%	74(45)	6(26)	13(54)	12(50)	12(52)	11(48)	11(46)	9(38)
04=>50%	40(24)	13(57)	5(21)	5(21)	1(4)	4(17)	5(21)	7(29)
CD34								
01=Dense	25(16)	4(18)	4(18)	4(18)	4(18)	3(14)	4(18)	2(10)
02=Moderately dense	57(37)	8(36)	7(32)	10(45)	6(27)	5(23)	9(41)	12(57)
03=Sparse	71(46)	10(45)	11(50)	8(36)	12(55)	14(64)	9(41)	7(33)
CD4								
01=0-5%	7(4)				1(5)	5(21)		1(4)
02=5-25%	54(33)	6(26)	7(29)	7(29)	8(36)	6(25)	9(38)	11(46)
03=25-50%	70(42)	8(35)	13(54)	10(42)	7(32)	10(42)	11(46)	11(46)
04=>50%	34(21)	9(39)	4(17)	7(29)	6(27)	3(13)	4(17)	1(4)
CD68_PCT								
01=0-10%	86(63)	11(46)	23(100)	19(83)	6(29)	9(39)		18(78)
02=10-20%	39(28)	10(42)		4(17)	13(62)	7(30)		5(22)
03=20-30%	9(7)	3(13)			2(10)	4(17)		
04=30-50%	3(2)					3(13)		
CD8								
01=0-5%	42(25)		2(9)	7(29)	18(75)	3(13)	3(13)	9(38)
02=5-25%	113(68)	18(78)	20(87)	16(67)	5(21)	20(83)	20(83)	14(58)
03=25-50%	4(2)	4(17)						
04=>50%	7(4)	1(4)	1(4)	1(4)	1(4)	1(4)	1(4)	1(4)
FOXP3_PCT,								
01=0-5%	96(62)	2(9)	13(57)	14(61)	16(73)	13(59)	19(83)	19(90)
02=5-25%	57(37)	19(86)	10(43)	9(39)	6(27)	7(32)	4(17)	2(10)
03=25-50%	3(2)	1(5)				2(9)		
Ki67								
01=0-5%	29(18)	5(22)	3(13)	5(22)	4(19)	11(46)		1(5)
02=5-25%	51(32)	14(61)	8(33)	3(13)	3(14)	10(42)	6(26)	7(32)
03=25-50%	57(36)	3(13)	8(33)	10(43)	11(52)	1(4)	14(61)	10(45)
04=>50%	23(14)	1(4)	5(21)	5(22)	3(14)	2(8)	3(13)	4(18)
CD3_ARCH								
01=Intrafollicular	3(2)			1(4)		1(4)		1(4)
02=Interfollicular	120(73)	21(91)	16(70)	10(42)	18(75)	12(52)	22(92)	21(88)
03=Diffuse	42(25)	2(9)	7(30)	13(54)	6(25)	10(43)	2(8)	2(8)
CD4_ARCH								
01=Intrafollicular	9(5)	1(4)	2(8)	1(4)	1(5)	1(4)	2(8)	1(4)
02=Interfollicular	118(71)	18(75)	18(75)	11(46)	18(82)	11(46)	19(79)	23(96)
03=Diffuse	39(23)	5(21)	4(17)	12(50)	3(14)	12(50)	3(13)	
CD68_ARCH								
01=Intrafollicular	15(9)	3(13)		3(13)	1(5)	3(13)	3(13)	2(8)
02=Interfollicular	74(45)	10(42)	12(50)	9(39)	6(29)	8(33)	13(54)	16(67)
03=Diffuse	75(46)	11(46)	12 (50)	11(48)	14 (67)	13(54)	8 (33)	6(25)
CD8_ARCH								
01=Intrafollicular	1(1)					1(4)		
02=Interfollicular	143(86)	22(92)	20(87)	20(83)	20(83)	16(67)	22(92)	23(96)
03=Diffuse	23(14)	2(8)	3(13)	4(17)	4(17)	7(29)	2(8)	1(4)
FOXP3_ARCH								
01=Intrafollicular	9(6)	2(9)			1(5)	4(18)		2(10)
02=Interfollicular	56(36)	9(41)	6(26)	9(41)	5(23)	8(36)	10(43)	9(43)
03=Diffuse	57(37)	7(32)	9(39)	8(36)	10(45)	10(45)	8(35)	5(24)
04=Perifollicular	33(21)	4(18)	8(35)	5(23)	6(27)		5(22)	5(24)

Table S3. Comparison of FOXP3, MIB1, CD68 scores from image analysis and manual (Core 1)

	Manual scoring category					Agreement measures for image and manual		Misclassification measure	
Method	NS*	0-5%	6-25%	26-50%	>50%	image = manual (%)	image within 5% manual (%)	image < manual (%)	image > manual (%)
FOXP3 Image		13	87	0	0				
Lab 1	9	13	70	9	0	81	95	50	50
Lab 2		57	43	0	0	48	96	8	92
Lab 3		48	52	0	0	57	96	10	90
Lab 4		74	26	0	0	30	91	6	94
Lab 5		57	30	13	0	35	91	27	73
Lab 6		78	22	0	0	35	91	0	100
Lab 7	4	87	9	0	0	23	95	0	100
MIB1 Image		39	61	4	0				
Lab 1	13	8	58	21		71	86	67	33
Lab 2	8	17	21	42	13	32	45	80	20
Lab 3	4	4	38	46	8	43	57	92	8
Lab 4	17	17	17	38	13	30	40	79	21
Lab 5	8	42	42	4	4	45	68	17	83
Lab 6		0	29	58	13	21	38	95	5
Lab 7	4	13	33	38	13	43	57	85	15
Method	NS	0-10%	11-20%	21-30%	31-50%				
CD68# Image		8	79	13	0				
Lab 1	4	38	33	21	4	61	87	78	22
Lab 2		100	0	0	0	38	75	0	100
Lab 3		79	21	0	0	46	88	8	92
Lab 4	13	29	38	21	0	52	86	70	30
Lab 5	4	33	25	17	21	17	52	68	32
Lab 7		75	25	0	0	54	88	9	91

*NS: not scored #Lab 6 did not score CD68

Table S4. Summary statistics for the comparison of image and manual scoring for FOXP3, MIB1 and CD68 for core 1. The median percent of cores and range within each category (as appropriate) is reported.

Method	Statistic	FOXP3	MIB1	CD68
Manual score distribution	Median percent of cores in each category across labs (range)			
	0-5% (0-10% for CD68)	57 (13-87)	13 (0-42)	51 (29-100)
	6-25% (11-20% for CD68)	30 (9-70)	33 (17-58)	25 (0-38)
	26-50% (30-50% for CD68)	0 (0-13)	38 (4-58)	8 (0-21)
	>50%	0	13 (4-13)	0 (0-21)
Categorized image scores by the manual categories	Percent of cores in 0-5%, 6-25%, 26-50% and >50%	13,87,0,0	39,61,4,0	8,79,13,0
Image vs. manual	Median percent of cores across labs (range)			
	Image=manual	35 (23-81)	43 (21-71)	49 (17-61)
	Image < manual	8 (0-50)	80 (17-95)	38 (0-78)
	Image>manual	92 (50-100)	20 (5-83)	62 (22-100)

Table S5. Reasons for not scoring samples by laboratories

	Lab							
Reason	1	2	3	4	5	6	7	Total
01=Non-representative core	5	1	4	21	1	0	5	37
02=Non-representative stain	5	2	1	4	0	1	0	13
03=Missing core	4	1	0	1	4	0	9	19
04=No internal control	1	0	0	0	1	0	0	2
05=Other	1	0	0	0	0	0	0	1
N/A	6	2	4	12	3	0	8	35
Total	22	6	9	38	9	1	22	107